
BIOSTATISTICAL ANALYSIS

FOURTH EDITION

JERROLD H. ZAR

Department of Biological Sciences
Northern Illinois University

PRENTICE HALL
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging in Publication Data

Zar, Jerrold H.
Biostatistical analysis / Jerrold H. Zar. — 4th ed.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 0-13-081542-X (alk. paper)
I. Biometry. I. Title
QH323.Z37 1999
S70'.1'5195—dc21

98-34062
CIP

Editorial/production supervision: *Interactive Composition Corporation*
Cover director: *Jayne Conte*
Cover designer: *Bruce Kenselaar*
Manufacturing manager: *Trudy Piscioti*
Editor: *Teresa Ryu*
Senior editor: *Sheri L. Snavely*
Editorial assistants: *Nancy Bauer and Lisa Tarabokja*

© 1999, 1996, 1984, 1974 by Prentice-Hall, Inc.
Upper Saddle River, New Jersey 07458

All rights reserved. No part of this book may be
reproduced, in any form or by any means,
without permission in writing from the publisher.

Printed in the United States of America

10 9

ISBN 0-13-081542-X

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S. A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

CONTENTS

PREFACE x	
1 INTRODUCTION 1	
1.1 Types of Biological Data 2	1.3 Frequency Distributions 6
1.2 Accuracy and Significant Figures 5	1.4 Cumulative Frequency Distributions 13
2 POPULATIONS AND SAMPLES 16	
2.1 Populations 16	2.3 Random Sampling 17
2.2 Samples from Populations 17	2.4 Parameters and Statistics 18
3 MEASURES OF CENTRAL TENDENCY 20	
3.1 The Arithmetic Mean 20	3.5 Other Measures of Central Tendency 28
3.2 The Median 23	3.6 The Effect of Coding Data 29
3.3 Other Quantiles 26	Exercises 31
3.4 The Mode 27	
4 MEASURES OF DISPERSION AND VARIABILITY 32	
4.1 The Range 32	4.6 The Coefficient of Variation 40
4.2 Dispersion Measured with Quantiles 34	4.7 Indices of Diversity 40
4.3 The Mean Deviation 34	4.8 The Effect of Coding Data 44
4.4 The Variance 35	Exercises 47
4.5 The Standard Deviation 39	
5 PROBABILITIES 48	
5.1 Counting Possible Outcomes 49	5.5 Probability of an Event 58
5.2 Permutations 50	5.6 Adding Probabilities 59
5.3 Combinations 54	5.7 Multiplying Probabilities 61
5.4 Sets 56	Exercises 63
6 THE NORMAL DISTRIBUTION 65	
6.1 Symmetry and Kurtosis 67	6.4 Introduction to Statistical Hypothesis Testing 79
6.2 Proportions of a Normal Distribution 72	6.5 Assessing Departures from Normality 86
6.3 The Distribution of Means 76	Exercises 89

- 7 ONE-SAMPLE HYPOTHESES 91**
- 7.1 Two-Tailed Hypotheses Concerning the Mean 91
 - 7.2 One-Tailed Hypotheses Concerning the Mean 96
 - 7.3 Confidence Limits for the Population Mean 98
 - 7.4 Reporting Variability about the Mean 100
 - 7.5 Sample Size and Estimation of the Population Mean 105
 - 7.6 Power and Sample Size in Tests Concerning the Mean 105
 - 7.7 Sampling Finite Populations 108
 - 7.8 Confidence Limits for the Population Median 110
 - 7.9 Hypotheses Concerning the Median 110
 - 7.10 Confidence Limits for the Population Variance 110
 - 7.11 Hypotheses Concerning the Variance 112
 - 7.12 Power and Sample Size in Tests Concerning the Variance 113
 - 7.13 Hypotheses Concerning the Coefficient of Variation 114
 - 7.14 Hypotheses Concerning Symmetry and Kurtosis 115
 - Exercises 120
- 8 TWO-SAMPLE HYPOTHESES 122**
- 8.1 Testing for Difference between Two Means 122
 - 8.2 Confidence Limits for Population Means 129
 - 8.3 Sample Size and Estimation of the Difference between Two Population Means 131
 - 8.4 Power and Sample Size in Tests for Difference between Two Means 132
 - 8.5 Testing for Difference between Two Variances 136
 - 8.6 Confidence Interval for the Population Variance Ratio 139
 - 8.7 Sample Size and Power in Tests for Difference between Two Variances 140
 - 8.8 Testing for Difference between Two Coefficients of Variation 141
 - 8.9 Nonparametric Statistical Methods 145
 - 8.10 Two-Sample Rank Testing 146
 - 8.11 Testing for Difference between Two Medians 155
 - 8.12 The Effect of Coding 155
 - 8.13 Two-Sample Testing of Nominal-Scale Data 156
 - 8.14 Testing for Difference between Two Diversity Indices 156
 - Exercises 159
- 9 PAIRED-SAMPLE HYPOTHESES 161**
- 9.1 Testing Mean Difference 161
 - 9.2 Confidence Limits for the Population Mean Difference 164
 - 9.3 Power and Sample Size in Paired-Sample Testing of Means 164
 - 9.4 Testing for the Difference between Variances from Two Correlated Populations 164
 - 9.5 Paired-Sample Testing by Ranks 165
 - 9.6 Confidence Limits for the Population Median Difference 169
 - 9.7 Paired-Sample Testing of Nominal-Scale Data 169
 - Exercises 175
- 10 MULTISAMPLE HYPOTHESES: THE ANALYSIS OF VARIANCE 177**
- 10.1 Single-Factor Analysis of Variance 178
 - 10.2 Confidence Limits for Population Means 189
 - 10.3 Power and Sample Size in Analysis of Variance 189
 - 10.4 Nonparametric Analysis of Variance 195
 - 10.5 Testing for Difference among Several Medians 200
 - 10.6 Homogeneity of Variances 202
 - 10.7 Homogeneity of Coefficients of Variation 204
 - 10.8 The Effect of Coding 206
 - 10.9 Multisample Testing for Nominal-Scale Data 206
 - Exercises 206
- 11 MULTIPLE COMPARISONS 208**
- 11.1 The Tukey Test 210
 - 11.2 The Newman-Keuls Test 214
 - 11.3 Confidence Intervals Following Multiple Comparisons 215
 - 11.4 Comparison of a Control Mean to Each Other Group Mean 217
 - 11.5 Scheffé's Multiple Contrasts 219
 - 11.6 Nonparametric Multiple Comparisons 223
 - 11.7 Nonparametric Multiple Contrasts 226
 - 11.8 Multiple Comparisons among Medians 226
 - 11.9 Multiple Comparisons among Variances 228
 - Exercises 230

Contents

12	TWO-FACTOR ANALYSIS OF VARIANCE	231
12.1	Two-Factor Analysis of Variance with Equal Replication	232
12.2	Two-Factor Analysis of Variance with Unequal Replication	245
12.3	Two-Factor Analysis of Variance without Replication	248
12.4	The Randomized Block Experimental Design	250
12.5	Repeated-Measures Experimental Designs	255
12.6	Multiple Comparisons and Confidence Intervals in Two-Factor Analysis of Variance	260
12.7	Power and Sample Size in Two-Factor Analysis of Variance	261
12.8	Nonparametric Randomized Block or Repeated-Measures Analysis of Variance	263
12.9	Multiple Comparisons for Nonparametric Randomized Block or Repeated-Measures Analysis of Variance	267
12.10	Dichotomous Nominal-Scale Data in Randomized Blocks or from Repeated Measures	268
12.11	Multiple Comparisons with Dichotomous Randomized Block or Repeated-Measures Data	270
12.12	Introduction to Analysis of Covariance	270
	Exercises	271
13	DATA TRANSFORMATIONS	273
13.1	The Logarithmic Transformation	275
13.2	The Square Root Transformation	275
13.3	The Arcsine Transformation	278
13.4	Other Transformations	280
	Exercises	280
14	MULTIWAY FACTORIAL ANALYSIS OF VARIANCE	282
14.1	Three-Factor Analysis of Variance	283
14.2	The Latin Square Experimental Design	286
14.3	Higher-Order Factorial Analysis of Variance	287
14.4	Blocked and Repeated-Measures Experimental Designs	288
14.5	Factorial Analysis of Variance with Unequal Replication	298
14.6	Multiple Comparisons and Confidence Intervals in Multiway Analysis of Variance	299
14.7	Power and Sample Size in Multiway Analysis of Variance	300
	Exercises	300
15	NESTED (HIERARCHICAL) ANALYSIS OF VARIANCE	303
15.1	Nesting within One Main Factor	305
15.2	Nesting in Factorial Experiments	308
15.3	Multiple Comparisons and Confidence Intervals	310
15.4	Power and Sample Size in Nested Analysis of Variance	311
	Exercise	311
16	MULTIVARIATE ANALYSIS OF VARIANCE	312
16.1	The Multivariate Normal Distribution	313
16.2	Multivariate Analysis of Variance Hypothesis Testing	316
16.3	Further Analysis	322
16.4	Other Experimental Designs	322
	Exercises	323
17	SIMPLE LINEAR REGRESSION	324
17.1	Regression vs. Correlation	324
17.2	The Simple Linear Regression Equation	326
17.3	Testing the Significance of a Regression	333
17.4	Confidence Intervals in Regression	337
17.5	Inverse Prediction	342
17.6	Interpretations of Regression Functions	344
17.7	Regression with Replication and Testing for Linearity	345
17.8	Power and Sample Size in Regression	350
17.9	Regression through the Origin	351
17.10	Data Transformations in Regression	353
17.11	The Effect of Coding	357
	Exercises	358

- 18 COMPARING SIMPLE LINEAR REGRESSION EQUATIONS 360**
- 18.1 Comparing Two Slopes 360
 - 18.2 Comparing Two Elevations 364
 - 18.3 Comparing Points on Two Regression Lines 368
 - 18.4 Comparing more than Two Slopes 369
 - 18.5 Comparing more than Two Elevations 372
 - 18.6 Multiple Comparisons among Slopes 372
 - 18.7 Multiple Comparisons among Elevations 373
 - 18.8 Multiple Comparisons of Points among Regression Lines 374
 - 18.9 An Overall Test for Coincidental Regressions 375
 - Exercises 375
- 19 SIMPLE LINEAR CORRELATION 377**
- 19.1 The Correlation Coefficient 377
 - 19.2 Hypotheses about the Correlation Coefficient 381
 - 19.3 Confidence Intervals for the Population Correlation Coefficient 383
 - 19.4 Power and Sample Size in Correlation 385
 - 19.5 Comparing Two Correlation Coefficients 386
 - 19.6 Power and Sample Size in Comparing Two Correlation Coefficients 388
 - 19.7 Comparing more than Two Correlation Coefficients 390
 - 19.8 Multiple Comparisons among Correlation Coefficients 392
 - 19.9 Rank Correlation 395
 - 19.10 Weighted Rank Correlation 398
 - 19.11 Correlation for Dichotomous Nominal-Scale Data 401
 - 19.12 Intraclass Correlation 404
 - 19.13 Concordance Correlation 407
 - 19.14 The Effect of Coding 410
 - Exercises 410
- 20 MULTIPLE REGRESSION AND CORRELATION 413**
- 20.1 Intermediate Computational Steps 414
 - 20.2 The Multiple Regression Equation 419
 - 20.3 Analysis of Variance of Multiple Regression or Correlation 422
 - 20.4 Hypotheses Concerning Partial Regression Coefficients 424
 - 20.5 Standardized Partial Regression Coefficients 426
 - 20.6 Partial Correlation 426
 - 20.7 Round-off Error and Coding Data 428
 - 20.8 Selection of Independent Variables 429
 - 20.9 Predicting Y Values 433
 - 20.10 Testing Difference between Two Partial Regression Coefficients 436
 - 20.11 "Dummy" Variables 436
 - 20.12 Interaction of Independent Variables 437
 - 20.13 Comparing Multiple Regression Equations 437
 - 20.14 Multiple Regression through the Origin 440
 - 20.15 Nonlinear Regression 440
 - 20.16 Descriptive vs. Predictive Models 442
 - 20.17 Concordance: Rank Correlation among Several Variables 443
 - Exercises 450
- 21 POLYNOMIAL REGRESSION 452**
- 21.1 Polynomial Curve Fitting 452
 - 21.2 Round-off Error and Coding Data 457
 - 21.3 Quadratic Regression 457
 - Exercises 459
- 22 TESTING FOR GOODNESS OF FIT 461**
- 22.1 Chi-Square Goodness of Fit 462
 - 22.2 Chi-Square Goodness of Fit for More than Two Categories 464
 - 22.3 Subdividing Chi-Square Analyses 466
 - 22.4 Chi-Square Correction for Continuity 468
 - 22.5 Bias in Chi-Square Calculations 470
 - 22.6 Heterogeneity Chi-Square 471
 - 22.7 The Log-Likelihood Ratio 473
 - 22.8 Kolmogorov-Smirnov Goodness of Fit for Discrete Data 475
 - 22.9 Kolmogorov-Smirnov Goodness of Fit for Continuous Data 478
 - 22.10 Sample Size Required for Kolmogorov-Smirnov Goodness of Fit for Continuous Data 481
 - Exercises 483

- 23 CONTINGENCY TABLES 486**
- 23.1 Chi-Square Analysis of Contingency Tables 488
- 23.2 Graphing Contingency Table Data 490
- 23.3 The 2×2 Contingency Table 491
- 23.4 Heterogeneity Testing of 2×2 Tables 500
- 23.5 Subdividing Contingency Tables 502
- 23.6 Bias in Chi-Square Contingency Table Analyses 504
- 23.7 The Log-Likelihood Ratio for Contingency Tables 505
- 23.8 Three-Dimensional Contingency Tables 506
- 23.9 Log-Linear Models for Multidimensional Contingency Tables 512
- Exercises 514
- 24 MORE ON DICHOTOMOUS VARIABLES 516**
- 24.1 Binomial Probabilities 517
- 24.2 The Hypergeometric Distribution 523
- 24.3 Sampling a Binomial Population 524
- 24.4 Confidence Limits for Population Proportions 527
- 24.5 Goodness of Fit for the Binomial Distribution 530
- 24.6 The Binomial Test 533
- 24.7 The Sign Test 538
- 24.8 Power of the Binomial and Sign Tests 539
- 24.9 Confidence Interval for the Population Median 542
- 24.10 The Fisher Exact Test 543
- 24.11 Comparing Two Proportions 555
- 24.12 Power and Sample Size in Comparing Two Proportions 558
- 24.13 Comparing more than Two Proportions 562
- 24.14 Multiple Comparisons for Proportions 563
- 24.15 Trends among Proportions 565
- Exercises 568
- 25 TESTING FOR RANDOMNESS 571**
- 25.1 Poisson Probabilities 571
- 25.2 Confidence Limits for the Poisson Parameter 574
- 25.3 Goodness of Fit of the Poisson Distribution 575
- 25.4 The Binomial Test Revisited 578
- 25.5 Comparing Two Poisson Counts 582
- 25.6 Serial Randomness of Nominal-Scale Categories 583
- 25.7 Serial Randomness of Measurements: Parametric Testing 586
- 25.8 Serial Randomness of Measurements: Nonparametric Testing 587
- Exercises 590
- 26 CIRCULAR DISTRIBUTIONS: DESCRIPTIVE STATISTICS 592**
- 26.1 Data on a Circular Scale 592
- 26.2 Graphical Presentation of Circular Data 595
- 26.3 Sines and Cosines of Circular Data 597
- 26.4 The Mean Angle 599
- 26.5 Angular Dispersion 602
- 26.6 The Median and Modal Angles 605
- 26.7 Confidence Limits for the Population Mean and Median Angles 605
- 26.8 Diametrically Bimodal Distributions 607
- 26.9 Second-Order Analysis: The Mean of Mean Angles 608
- 26.10 Confidence Limits for the Second-Order Mean Angle 611
- Exercises 614
- 27 CIRCULAR DISTRIBUTIONS: HYPOTHESIS TESTING 616**
- 27.1 Testing Significance of the Mean Angle: Unimodal Distributions 616
- 27.2 Testing Significance of the Median Angle: Omnibus Test 621
- 27.3 Testing Significance of the Median Angle: Binomial Test 624
- 27.4 Testing Symmetry around the Median Angle 624
- 27.5 Two-Sample and Multisample Testing of Mean Angles 625
- 27.6 Nonparametric Two-Sample and Multisample Testing of Angles 630
- 27.7 Two-Sample and Multisample Testing of Median Angles 635
- 27.8 Two-Sample and Multisample Testing of Angular Distances 635
- 27.9 Two-Sample and Multisample Testing of Angular Dispersion 637
- 27.10 Parametric One-Sample Second-Order Analysis of Angles 638

27 CIRCULAR DISTRIBUTIONS: HYPOTHESIS TESTING (continued)

- | | |
|--|---|
| 27.11 Nonparametric One-Sample Second-Order Analysis of Angles 639 | 27.16 Parametric Angular Correlation and Regression 649 |
| 27.12 Parametric Two-Sample Second-Order Analysis of Angles 641 | 27.17 Nonparametric Angular Correlation 653 |
| 27.13 Nonparametric Two-Sample Second-Order Analysis of Angles 643 | 27.18 Goodness of Fit Testing for Circular Distributions 654 |
| 27.14 Parametric Paired-Sample Testing with Angles 645 | 27.19 Serial Randomness of Nominal-Scale Categories on a Circle 658 |
| 27.15 Nonparametric Paired-Sample Testing with Angles 647 | Exercises 660 |

APPENDIX A ANALYSIS OF VARIANCE HYPOTHESIS TESTING App1

- | | |
|---|--|
| A.1 Determination of Appropriate F 's and Degrees of Freedom App1 | A.4 Nested Analysis of Variance App7 |
| A.2 Two-Factor Analysis of Variance App5 | A.5 Split-Plot and Mixed Within-Subjects Analysis of Variance App8 |
| A.3 Three-Factor Analysis of Variance App6 | |

APPENDIX B STATISTICAL TABLES AND GRAPHS App11

- | | |
|--|--|
| Table B.1 Critical Values of Chi-Square Distribution App12 | Table B.16 Critical Values of Q' for Nonparametric Multiple Comparison Testing with a Control App108 |
| Table B.2 Proportions of the Normal Curve (One-Tailed) App17 | Table B.17 Critical Values of the Correlation Coefficient, r App109 |
| Table B.3 Critical Values of the t Distribution App19 | Table B.18 Fisher's z Transformation for Correlation Coefficients, r App111 |
| Table B.4 Critical Values of the F Distribution App21 | Table B.19 Correlation Coefficients, r , Corresponding to Fisher's z Transformation App113 |
| Table B.5 Critical Values of the q Distribution App58 | Table B.20 Critical Values of the Spearman Rank Correlation Coefficient, r_s App116 |
| Table B.6 Critical Values of q' for the One-Tailed Dunnett's Test App74 | Table B.21 Critical Values of the Top-Down Correlation Coefficient, r_T App118 |
| Table B.7 Critical Values of q' for the Two-Tailed Dunnett's Test App76 | Table B.22 Critical Values of the Symmetry Measure, g_1 App119 |
| Table B.8 Critical Values of d_{max} for the Kolmogorov-Smirnov Goodness of Fit Test for Discrete or Grouped Data App77 | Table B.23 Critical Values of the Kurtosis Measure, g_2 App121 |
| Table B.9 Critical Values of D for the Kolmogorov-Smirnov Goodness of Fit Test for Continuous Distributions App83 | Table B.24 The Arcsine Transformation, p' App124 |
| Table B.10 Critical Values of D_s for the δ -Corrected Kolmogorov-Smirnov Goodness of Fit Test for Continuous Distributions App87 | Table B.25 Proportions, p , Corresponding to Arcsine Transformations, p' App127 |
| Table B.11 Critical Values of the Mann-Whitney U Distribution App89 | Table B.26a Binomial Coefficients, ${}_nC_X$ App129 |
| Table B.12 Critical Values of the Wilcoxon T Distribution App101 | Table B.26b Proportions of the Binomial Distribution for $p = q = 0.5$ App132 |
| Table B.13 Critical Values of the Kruskal-Wallis H Distribution App104 | Table B.27 Critical Values of C for the Sign Test or for the Binomial Test with $p = 0.5$ App133 |
| Table B.14 Critical Values of the Friedman χ^2 Distribution App106 | Table B.28 Critical Values for Fisher's Exact Test App143 |
| Table B.15 Critical Values of Q for Nonparametric Multiple Comparison Testing App107 | Table B.29 Critical Values for Runs Test App171 |
| | Table B.30 Critical Values of C for the Mean Square Successive Difference Test App180 |
| | Table B.31 Critical Values for the Runs Up and Down Test App182 |

APPENDIX B STATISTICAL TABLES AND GRAPHS (continued)

Table B.32 Angular Deviation, s , As a Function of
Vector Length, r App184

Table B.33 Circular Standard Deviation, s_0 , As a
Function of Vector Length, r App186

Table B.34 Critical Values of Rayleigh's z App188

Table B.35 Critical Values of u for the V Test of
Circular Uniformity App190

Table B.36 Critical Values of m for the Hodges-Ajne
Test App191

Table B.37 Correction Factor, K , for the Watson
and Williams Test App193

Table B.38 Critical Values of Watson's U^2 App195

Table B.39 Critical Values of R' for the Moore Test
of Circular Uniformity App198

Table B.40 Common Logarithms of Factorials App199

Table B.41 Ten Thousand Random Digits App201

Figure B.1 Power and Sample Size in Analysis of
Variance App205

ANSWERS TO EXERCISES *Ans1*

LITERATURE CITED *L1*

INDEX *I1*

would be described by the *binomial distribution* (sometimes referred to as the "Bernoulli distribution"). Let us now examine binomial probabilities.

24.1 BINOMIAL PROBABILITIES

Consider a population consisting of two categories, where p is the proportion of individuals in one of the categories and $q = 1 - p$ is the proportion in the other. Then the probability of selecting at random from this population a member of the first category is p , and the probability of selecting a member of the second category is q .[†]

For example, let us say we have a population of female and male animals, in proportions of $p = 0.4$ and $q = 0.6$, respectively, and we take a random sample of two individuals from the population. The probability of the first being a female is p (i.e., 0.4) and the probability of the second being a female is also p . As the probability of two mutually exclusive events both occurring is the product of the probabilities of the two separate events (Section 5.7), the probability of having two females in a sample of two is $(p)(p) = p^2 = 0.16$; the probability of the sample of two consisting of two males is $(q)(q) = q^2 = 0.36$.

What is the probability of the sample of two consisting of one male and one female? This could occur by the first individual being a female and the second a male (with a probability of pq) or by the first being a male and the second a female (which would occur with a probability of qp). The probability of either of two mutually exclusive outcomes is the sum of the probabilities of each outcome (Section 5.6), so the probability of one female and one male in the sample is $pq + qp = 2pq = 2(0.4)(0.6) = 0.48$. Note that $0.16 + 0.36 + 0.48 = 1.00$.

Now consider another sample from this population, one where $n = 3$. The probability of all three individuals being female is $ppp = p^3 = (0.4)^3 = 0.064$. The probability of two females and one male is ppq (for a sequence of ♀ ♀ ♂) + pqp (for ♀ ♂ ♀) + qpp (for ♂ ♀ ♀), or $3p^2q = 3(0.4)^2(0.6) = 0.288$. The probability of one female and two males is pqq (for ♀ ♂ ♂) + qpq (for ♂ ♀ ♂) + qqp (for ♂ ♂ ♀), or $3pq^2 = 3(0.4)(0.6)^2 = 0.432$. And, finally, the probability of all three being males is $qqq = q^3 = (0.6)^3 = 0.216$. Note that $p^3 + 3p^2q + 3pq^2 + q^3 = 0.064 + 0.288 + 0.432 + 0.216 = 1.000$ (meaning that there is a 100% probability—that is, it is certain—that the three animals will be in one of these three combinations of sexes).

*The binomial formula discussed in the following section was first described in 1676 by Sir Isaac Newton (1642–1727), the great English scientist and mathematician; and its first proof, for positive integer exponents, was given by the Swiss mathematician, Jacques (also known as Jakob or James) Bernoulli (1654–1705) in a 1713 publication (Cajori, 1954). Each observed event from a binomial distribution is sometimes called a "Bernoulli trial." David (1995) ascribes the first use of the term "binomial distribution" to G. U. Yule, in 1911.

†This assumes "sampling with replacement." That is, each individual in the sample is taken at random from the population and then is returned to the population before the next member of the sample is selected. Sampling without replacement is discussed in Section 24.2. If the population is very large compared to the size of the sample, then sampling with and without replacement are indistinguishable in practice.

If we performed the same exercise with $n = 4$, we would find that the probability of four females is $p^4 = (0.4)^4 = 0.0256$, the probability of three females (and one male) is $4p^3q = 4(0.4)^3(0.6) = 0.1536$, the probability of two females is $6p^2q^2 = 0.3456$, the probability of one female is $4pq^3 = 0.3456$, and the probability of no females (i.e., all four are male) is $q^4 = 0.1296$. (The sum of these five terms is 1.0000, a good arithmetic check.)

If a random sample of size n is taken from a binomial population, then the probability of X individuals being in one category (and, therefore, $n - X$ individuals in the second category) is

$$P(X) = \binom{n}{X} p^X q^{n-X}. \quad (24.1)$$

In this equation, $p^X q^{n-X}$ refers to the probability of sample consisting of X items, each having a probability of p , and $n - X$ items, each with probability q . The *binomial coefficient*,

$$\binom{n}{X} = \frac{n!}{X!(n-X)!}, \quad (24.2)$$

is the number of ways X items of one kind can be arranged with $n - X$ items of a second kind, or, in other words, the number of possible *combinations* of n items divided into one group of X items and a second group of $n - X$ items. (See Section 5.3 for a discussion of combinations; Equation 5.3 explained the factorial notation, "!"). Therefore, Equation 24.1 can be written as

$$P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}. \quad (24.3)$$

Thus, $\binom{n}{X} p^X q^{n-X}$ is the X th term in the expansion of $(p + q)^n$, and Table 24.1 shows this expansion for powers up through 6. Note that for any power, n , the sum of the two exponents in any term is n . Furthermore, the first term will always be p^n , the second will always contain $p^{n-1}q$, the third will always contain $p^{n-2}q^2$, etc., with the last term always being q^n . The sum of all the terms in a binomial expansion will always be 1.0, for $p + q = 1$, and $(p + q)^n = 1^n = 1$.

As for the coefficients of these terms in the binomial expansion, the X th term of the n th power expansion can be calculated by Equation 24.3. Furthermore, the examination of these coefficients as shown in Table 24.2 has been deemed interesting for centuries.

TABLE 24.1 Expansion of the Binomial, $(p + q)^n$

n	$(p + q)^n$
1	$p + q$
2	$p^2 + 2pq + q^2$
3	$p^3 + 3p^2q + 3pq^2 + q^3$
4	$p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4$
5	$p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$
6	$p^6 + 6p^5q + 15p^4q^2 + 20p^3q^3 + 15p^2q^4 + 6pq^5 + q^6$

TABLE 24.2 Binomial Coefficient, ${}_n C_X$

n	$X=0$	1	2	3	4	5	6	7	8	9	10	Sum of coefficients
1	1	1										$2 = 2^1$
2	1	2	1									$4 = 2^2$
3	1	3	3	1								$8 = 2^3$
4	1	4	6	4	1							$16 = 2^4$
5	1	5	10	10	5	1						$32 = 2^5$
6	1	6	15	20	15	6	1					$64 = 2^6$
7	1	7	21	35	35	21	7	1				$128 = 2^7$
8	1	8	28	56	70	56	28	8	1			$256 = 2^8$
9	1	9	36	84	126	126	84	36	9	1		$512 = 2^9$
10	1	10	45	120	210	252	210	120	45	10	1	$1024 = 2^{10}$

This arrangement is known as *Pascal's triangle*.^{*} We can see from this triangular array that any binomial coefficient is the sum of two coefficients on the line above it, namely,

$$\binom{n}{X} = \binom{n-1}{X-1} + \binom{n-1}{X}. \quad (24.4)$$

This can be more readily observed if we display the triangular array as follows:

				1										
				1		1								
				1		2		1						
				1		3		3		1				
				1		4		6		4		1		
				1		5		10		10		5		1

Also note that the sum of all coefficients for the n th power binomial expansion is 2^n . Appendix Table B.26a presents binomial coefficients for much larger n 's and X 's, and they will be found useful later in this chapter.

Thus, we can calculate probabilities of category frequencies occurring in random samples from binomial population. If, for example, a sample of five (i.e., $n = 5$) is taken from a population composed of 50% males and 50% females (i.e., $p = 0.5$ and $q = 0.5$) then Example 24.1 shows how Equation 24.3 is used to determine the probability of the sample containing 0 males, 1 male, 2 males, 3 males, 4 males, and 5 males. These

^{*}Blaise Pascal (1623–1662), French mathematician and physicist and one of the founders of probability theory (in 1654, immediately before abandoning mathematics to become a religious recluse). He had his triangular binomial coefficient derivation published in 1665, although knowledge of the triangular properties appears in Chinese writings as early as 1303 (Cajori, 1954; David, 1962; Struik, 1967: 79). Pascal also invented (at age 19) a mechanical adding and subtracting machine which, though patented in 1649, proved too expensive to be practical to construct (Asimov, 1982: 130–131). His significant contributions to the study of fluid pressures have been honored by naming the international unit of pressure the pascal, which is a pressure of one newton per square meter (where a newton—named for Sir Isaac Newton—is the unit of force representing a one-kilogram mass accelerating at the rate of one meter per second per second). Pascal is also the name given to a modern computer language. The relationship of Pascal's triangle to ${}_n C_X$ was first published in 1685 by the English mathematician, John Wallis (1616–1703) (David, 1962: 123–124).

probabilities are found to be 0.03125, 0.15625, 0.31250, 0.31250, 0.15625, and 0.03125, respectively. This enables us to state that if we took 100 random samples of five animals each from the population, about three of the sample [i.e., $(0.03125)(100) = 3.125$ of them] would be expected to contain all females, about sixteen [i.e., $(0.15625)(100) = 15.625$] to contain one male and four females, thirty-one [i.e., $(0.31250)(100)$] to consist of two males and three females, etc. If we took 1400 random samples of five, then $(0.03125)(1400) = 43.75$ [i.e., about 44] of them would be expected to contain all females, etc. Figure 24.1a shows graphically the binomial distribution for $p = q = 0.5$, for $n = 5$. Note, from Fig 24.1a and Example 24.1, that when $p = q = 0.5$ the distribution is symmetrical [i.e., $P(0) = P(5)$, $P(1) = P(4)$, etc.], and Equation 24.3 becomes

$$P(X) = \frac{n!}{X!(n-X)!} 0.5^n. \quad (24.5)$$

Appendix Table B.26b gives binomial probabilities for $n = 2$ to $n = 20$, for $p = 0.5$.

Example 24.2 presents the calculation of binomial probabilities for the case where $n = 5$, $p = 0.3$, and $q = 1 - 0.3 = 0.7$. Thus, if one were sampling a population consisting of 30% males and 70% females, 0.16807 (i.e., 16.807%) of the samples would be expected to contain no males, 0.36015 to contain one male and four females, etc. Fig. 24.1b presents this binomial distribution graphically, whereas Fig. 24.1c shows the distribution where $p = 0.1$ and $q = 0.9$.

For calculating binomial probabilities for large n , it is often convenient to employ logarithms. For this reason, Appendix Table B.40, a table of logarithms of factorials, is provided. Alternatively, it is useful to note that the denominator of Equation 24.3 cancels out much of the numerator, so that it is possible to simplify the computation of $P(X)$,

EXAMPLE 24.1 Computing binomial probabilities, $P(X)$, where $n = 5$, $p = 0.5$, and $q = 0.5$ (following Equation 24.3).

X	$P(X)$
0	$\frac{5!}{0!5!} (0.5^0)(0.5^5) = (1)(1.0)(0.03125) = 0.03125$
1	$\frac{5!}{1!4!} (0.5^1)(0.5^4) = (5)(0.5)(0.0625) = 0.15625$
2	$\frac{5!}{2!3!} (0.5^2)(0.5^3) = (10)(0.25)(0.125) = 0.31250$
3	$\frac{5!}{3!2!} (0.5^3)(0.5^2) = (10)(0.125)(0.25) = 0.31250$
4	$\frac{5!}{4!1!} (0.5^4)(0.5^1) = (5)(0.0625)(0.5) = 0.15625$
5	$\frac{5!}{5!0!} (0.5^5)(0.5^0) = (1)(0.03125)(1.0) = 0.03125$

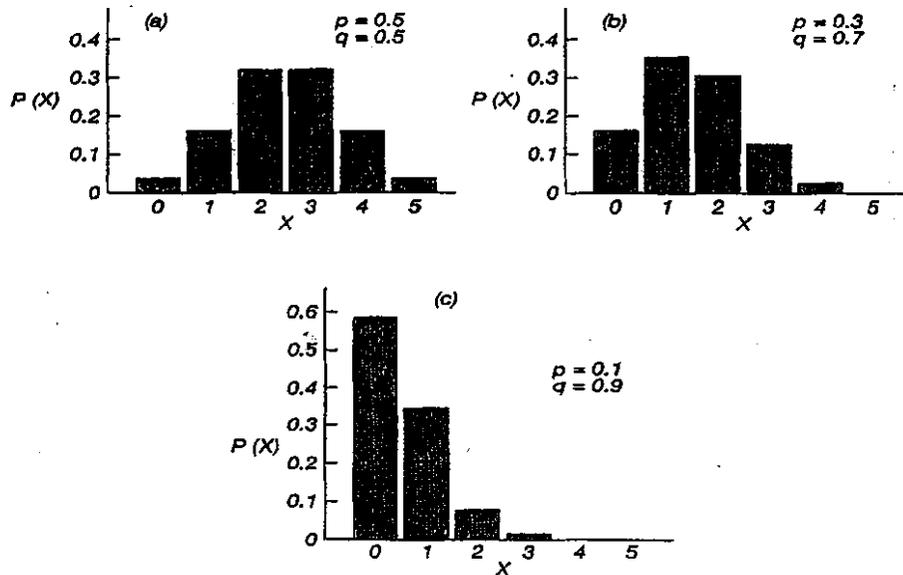


Figure 24.1 The binomial distribution, for $n = 5$. (a) $p = q = 0.5$. (b) $p = 0.3$, $q = 0.7$. (c) $p = 0.1$, $q = 0.9$. These graphs were drawn utilizing the proportions given by Equation 24.1.

especially in the tails of the distribution (i.e., for low X and for high X), as shown in Example 24.3. If p is very small, then the use of the Poisson distribution (Section 25.1), should be considered.*

The mean of a binomial distribution of counts X , is

$$\mu_X = np, \quad (24.6)$$

the variance† is

$$\sigma_X^2 = npq, \quad (24.8)$$

and the standard deviation of X is

$$\sigma_X = \sqrt{npq}. \quad (24.9)$$

*Raff (1956) and Molenaar (1969a, 1969b) discuss several approximations to the binomial distribution, including the normal and Poisson distributions.

†A measure of symmetry (see Section 6.1) for a binomial distribution is

$$\gamma_1 = \frac{q-p}{\sqrt{npq}}. \quad (24.7)$$

so it can be seen that $\gamma_1 = 0$ only when $p = q = 0.5$, $\gamma_1 > 0$ implies a distribution skewed to the right (as in Figs. 24.1b and 24.1c) and $\gamma_1 < 0$ indicates a distribution skewed to the left.

EXAMPLE 24.2 Computing binomial probabilities, $P(X)$, where $n = 5$, $p = 0.4$, $q = 0.7$ (following Equation 24.3).

X	$P(X)$
0	$\frac{5!}{0!5!} (0.3^0)(0.7^5) = (1)(1.0)(0.16807) = 0.16807$
1	$\frac{5!}{1!4!} (0.3^1)(0.7^4) = (5)(0.3)(0.2401) = 0.36015$
2	$\frac{5!}{2!3!} (0.3^2)(0.7^3) = (10)(0.09)(0.343) = 0.30870$
3	$\frac{5!}{3!2!} (0.3^3)(0.7^2) = (10)(0.027)(0.49) = 0.13230$
4	$\frac{5!}{4!1!} (0.3^4)(0.7^1) = (5)(0.0081)(0.7) = 0.02835$
5	$\frac{5!}{5!0!} (0.3^5)(0.7^0) = (1)(0.00243)(1.0) = 0.00243$

EXAMPLE 24.3 Computing binomial probabilities, $P(X)$, with $n = 400$, $p = 0.02$, and $q = 0.98$.

(Many calculators can operate with large powers of numbers; otherwise, logarithms may be used.)

X	$P(X)$
0	$\frac{n!}{0!(n-0)!} p^0 q^{n-0} = q^n = 0.98^{400} = 0.00031$
1	$\frac{n!}{1!(n-1)!} p^1 q^{n-1} = npq^{n-1} = (400)(0.02)(0.98^{399}) = 0.00253$
2	$\frac{n!}{2!(n-2)!} p^2 q^{n-2} = \frac{n(n-1)}{2!} p^2 q^{n-2} = \frac{(400)(399)}{2} (0.02^2)(0.98^{398}) = 0.01028$
3	$\frac{n!}{3!(n-3)!} p^3 q^{n-3} = \frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} = \frac{(400)(399)(398)}{(3)(2)} (0.02^3)(0.98^{397}) = 0.02784$

and so on.

Thus, if we have a binomially distributed population where p (e.g., the proportion of males) = 0.5 and q (e.g., the proportion of females) = 0.5 and we take ten samples from that population, the mean of the ten X 's (i.e., the mean number of males per sample) would be expected to be $np = (10)(0.05) = 5$ and the standard deviation of the ten X 's would be expected to be $\sqrt{npq} = \sqrt{(10)(0.5)(0.5)} = 1.58$. Our concern typically is with the distribution of the expected probabilities rather than the expected X 's, as will be explained in Section 24.3.

